WP. 51
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

Data Utility, Confidentiality, and the Production-Possibility Frontier: Striking a Delicate Balance

Prepared by Daniel Beckler, USDA – National Agricultural Statistics Service and Timothy Mulcahy, NORC at the University of Chicago, U.S.A.

Data Utility, Confidentiality, and the Production-Possibility Frontier: Striking a Delicate Balance

Daniel Beckler* and Timothy Mulcahy**

* USDA – National Agricultural Statistics Service

** NORC at the University of Chicago

1 Introduction

The United States Department of Agriculture's National Agricultural Statistics Service (USDA NASS) conducts hundreds of surveys annually and the Census of Agriculture every five years on the nation's farmers and agribusinesses. These data collection efforts provide the basic data from which official USDA estimates are derived for virtually every facet of United States (US) agriculture. In addition, some of the census and survey datasets provide rich analytic utility to both academic and governmental research entities. NASS, working in partnership with the USDA's Economic Research Service (ERS), and NORC at the University of Chicago, have implemented methods to ensure the confidentiality of the record-level census and survey data while providing researcher access to data and tools to improve the agricultural industry and guide agricultural policy.

This paper will provide an overview of various modalities available to make sensitive data available to approved researchers. The paper also will provide an overview of NASS and its approaches to data accessibility and maintaining confidentiality, while providing maximum data utility to its customers.

2 Striking a Delicate Balance between Confidentiality and Utility

There are numerous ways in which data producers can disseminate microdata. Ultimately, in determining the most appropriate data access modality, data producers must examine the tradeoffs between confidentiality, analytic utility, and convenience to access (Lane, Heus, & Mulcahy, 2008). While one's primary consideration is guaranteeing a high level of data security and data confidentiality as appropriate to the sensitivity of the data, the data producer must also ensure assure that the files are analytically valid. The resulting dataset(s) also must be analytically meaningful and actually used by the intended audience(s) (Winkler, 1997).

Datasets can be statistically perturbed to reduce risk of disclosure and protect data confidentiality. This however comes at a cost to data quality and analytic utility. Custodians of data therefore must strike a delicate balance between protecting confidentiality and maximizing analytic utility. Given that data providers cannot guarantee “zero disclosure risk”, the focus is squarely on reducing risk to its lowest possible point while maximizing data quality for analytic utility. Next, we provide an overview of current data access modality options.

2.1 Public Use Files

Data producers often release microdata via public use files (PUFs) through the Internet, thus providing public access to datasets that have undergone statistical protection techniques such as variable level suppression, top and bottom coding, noise infusion, geographic aggregation, etc. to protect data confidentiality (Weinberg et al., 2007). While such techniques are necessary, statistically treating the data by definition diminishes data quality (United Nations, 2007). Still PUFs serve an important role in today’s society, for example in training student researchers. Indeed PUFs are a safe means by which student researchers may gain proficiency in conceptualizing research questions and conducting analyses on safe datasets before moving on to using sensitive variables contained in untreated, “raw” datasets.

2.2 Online Statistical Data Cubes and Tabulation Engines

Online statistical data cubes and tabulation engines provide researchers controlled access to raw microdata. Authorized users submit data queries online and receive output in real time. Output derived from data cubes generally is as safe as PUFs in that all results emanate from pre-defined, tabular outputs. Every possible combination of queries and derivative output (i.e., tables) that is contained in the backend database, a priori, has been rendered safe. Similarly, online tabulation engines allow researchers to pose questions of their own choosing and execute queries via the Internet, filtering on sensitive variables. Data confidentiality in this scenario is protected in that researcher output is returned in the form of customized summary tables that have undergone automated disclosure control treatment– either using variable level suppression or statistical masking techniques.

2.3 Remote Batch Processing

Remote batch processing offers another useful dissemination modality. Rather than providing researchers access to raw datasets, users submit programs or code remotely via the Internet or email. Output subsequently is returned to researchers after undergoing statistical disclosure control processing. After output (i.e., results) is determined as safe, either by data disclosure analysts or by an automatic statistical disclosure control (SDC) procedure, it may be emailed to researchers. While most batch processing systems use filters or algorithms to suppress certain queries and results, the output is generally of greater analytic value than that which is obtained using PUFs (Weinberg et al, 2007). What’s more, while batch execution processing jobs are relatively secure and effective for simple requests, comparatively more complex queries require significantly more computational power and often result in slowness in returning output (United Nations, 2007).

A general theme emerges from the aforementioned dissemination modalities, i.e., there are serious tradeoffs that must be considered before selecting the most appropriate data access solution, including data utility, confidentiality, security, and ease of use (i.e., convenience). While PUFs, remote batch processing, and tabulation engines are easy to access and may have imbedded security measures to protect data confidentiality, by definition these measures reduce analytic utility. There are, however, other options available to data providers that allow researchers to increase data analytical utility.

2.4 Synthetic Microdata

One such example is the development and use of synthetic microdata, i.e., data that are simulated to reproduce the statistical properties of the underlying confidential data. In this scenario, not only is data confidentiality preserved but so too is analytic quality. Although the full benefit of developing synthetic data depends on the validity of the models that have been used to create it (Schueren, 2009); provided that the models are accurate, the estimates and inferences derived from the treated data will be a very good fit to the original, untreated dataset.

2.5 Remote and Physical Data Enclaves

Remote and physical data enclaves (also referred to as research data centers (RDCs) offer high levels of security. Whereas remote access platforms provide convenient access via an encrypted terminal session, RDCs typically provide on-site access only. To protect confidentiality, remote and physical data enclaves maintain stringent physical and computer security guidelines. In addition, all output (i.e., results) that researchers request for export undergoes a formal disclosure review process before it is exported from the controlled environment and hence made public.

An obvious advantage of remote and physical data enclaves is that researchers oftentimes have access to the most detailed version of the data, i.e., raw microdata, devoid of statistical treatment (data perturbation, suppression, etc.). Access to such analytically useful data through RDCs however, also comes at a price, in that they are very expensive to operate and sometimes inconvenient for researchers. Access via RDCs also requires researchers to be physically present at the facility and the process for reviewing proposals or what results may be publicly released out of the RDC is reportedly very cumbersome and time-consuming (United Nations, 2007).

2.6 Technical Solution

Often using virtual private network (VPN) technology to provide remote data access, remote data enclaves typically allow approved researchers to connect to a data server that hosts the actual microdata and work in a remote-desktop environment (virtual machines). While users work in a familiar environment (Windows, Linux, etc), and have access to a full suite of statistical applications, no output may leave the secured environment without first undergoing stringent statistical disclosure control.

Of all the aforementioned modalities, in terms of disclosure risk, data's analytical utility, and ease of access, remote data enclaves achieve the optimal production frontier that meets all three objectives. A visual representation of different modalities' strength in maintaining confidentiality and providing data with analytic utility is shown below in Figure 1.

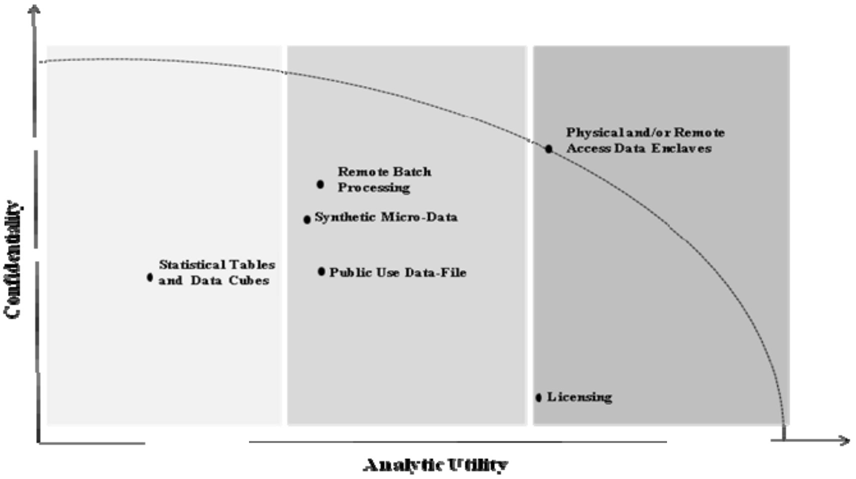

Figure 1

In terms of convenience, remote enclave users are no longer burdened with the need to travel to physical data centers. Indeed, they may access the microdata anytime at their own convenience. Data enclaves therefore offer high security, high analytic quality, and high convenience. The additional value is depicted below in Figure 2.

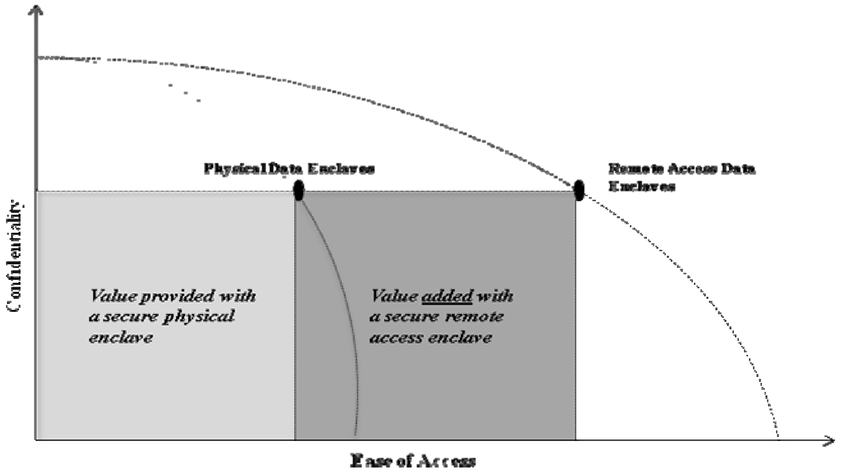
Confidentiality
Physical Data Enclaves
Remote Access Data Enclaves
Value provided with a secure physical enclave
Value added with a secure remote access enclave
Ease of Access

Figure 2:

3 An Overview of the National Agricultural Statistics Service

NASS's primary mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. NASS accomplishes its mission by conducting voluntary sample surveys and the mandatory Census of Agriculture. Depending on their purposes, the sample surveys are conducted on fixed intervals that range from weekly to annually. The Census of Agriculture is conducted every five years (in years ending with a "2" or "7") and provides detailed data at the county level. The official estimates generated from the sample surveys are closely watched by the agricultural industry and influence commodity markets worldwide; some are Primary Economic Indicators of the United States.[1]

3.1 NASS's Customers

NASS's customer base is diverse and includes farmers, agricultural associations, commodity traders, academia, commercial firms, and governmental policy makers (Osborn & Barr, 2011). Each of these entities has its own needs and sophistication level. Academics and policy makers tend to be more interested in long-term trends. Farmers, agricultural associations and commodity traders tend to be interested in only the most current official estimates as their business decisions are most affected by the current snapshot of the agricultural industry. These data users are [anecdotally] less concerned with data quality metrics associated with the estimates, as the agricultural industry and commodity markets react to the official estimates NASS publishes regardless of their inherent quality.

3.2 Census of Agriculture Suppression Methodology for Confidentiality

The suppression methodology for confidentiality used for the Census of Agriculture (COA) aims to protect published totals from revealing an individual agricultural operation's information or allow it to be closely estimated. Estimates deemed as sensitive are suppressed in publications made available to the public. Farm and operator counts are not considered to be sensitive, and are never suppressed.

The specific COA methodology for identifying sensitive totals is a combination of a threshold rule and a dominance rule. The threshold rule requires a minimum number of positive values that contribute to the total; the dominance rule is the classic (n,k) rule. For the threshold rule, totals are considered sensitive, and thus suppressed, if fewer than three positive reports contribute to the total. For the dominance rule,

[1] Data provided by census and survey respondents are protected from both individual and aggregate disclosure by Title 7, Section 2276 of the United States Code and by Title 5, Confidential Information Protection and Statistical Efficiency (CIPSEA). All NASS staff must annually sign a *Confidentiality Certification* that identifies all laws and NASS policies that protect respondent-reported and other unpublished data from disclosure to the public.

totals are considered sensitive and suppressed if the largest *n* positive respondents that contribute to the total make up at least *k* percent of that total. The specific values of *n* and *k* are not revealed to the public, as their disclosure would lessen the effectiveness of the suppressions.

Suppressions resulting from the threshold and dominance rules are referred to as *primary suppressions*. In addition to the primary suppressions, a computer routine using a Minimum Cost Flow (MCF) algorithm identifies *complementary suppressions* that are needed to prevent primarily suppressed values from being calculated from linear combinations of non-suppressed estimates. The complementary suppressions are chosen such to minimize a cost function, where the cost is the relative size of the candidate values in the cells making up the linear combinations of non-suppressed estimates. In order to maximize the utility of the COA, the number of complementary suppressions is desired to be minimized. Although the MCF algorithm identifies a minimum number of complementary suppressions in any given published table, it unfortunately does not ensure an optimal solution across multiple, related tables. This lack of optimality reduces the overall data utility of the COA by possibly creating more complementary suppressions than necessary across tables.

After primary and complementary suppressions are made, NASS analysts review the final tables prior to publication and have the opportunity to override the computer-selected complementary suppressions and choose alternative values to suppress. This manual review increases the overall data utility of the COA since it ensures the computer-selected complementary suppressions do not withhold estimates deemed vital to NASS's data users.

One of the main COA data products is a series of US and state-level publications that contain numerous cross-tabulation tables. Table 1 provides the counts of the number of possible non-zero estimates of totals published in all tables contained in the US and state-level publications for the 2007 COA. Also included are the number of suppressions (primary, complementary, and total) and the percentage of all possible estimates that were suppressed. The table shows that the number of suppressions made to ensure confidentiality was quite small – 2.08 percent – for the US totals, and ranged from a high of 45.39 percent to a low of 18.58 percent at the state level. Considering all possible estimates across both US and state-level publications, 22.78 percent were suppressed. Although the number of state-level suppressions was sometimes quite large, Table 1 further shows that primary suppressions dominated the total, with nearly 74 percent (430,843) of the overall suppressions being primary.

Domain	Overall Count of Estimates 1/	Number of Primary Suppressions	Number of Complimentary Suppressions	Total Number of Suppressions	Total Suppressions as Percent of Count of Estimates
United States	29,075	255	351	606	2.08
Alabama	54,438	9,437	3,077	12,514	22.99
Alaska	16,095	5,111	2,195	7,306	45.39
Arizona	29,476	5,554	2,737	8,291	28.13
Arkansas	54,712	8,878	2,842	11,720	21.42
California	60,472	9,456	3,466	12,922	21.37
Colorado	49,949	7,807	2,614	10,421	20.86
Connecticut	23,041	3,460	2,747	6,207	26.94
Delaware	20,533	4,325	2,462	6,787	33.05
Florida	55,266	10,708	3,633	14,341	25.95
Georgia	89,587	18,345	5,190	23,535	26.27
Hawaii	18,500	3,334	2,331	5,665	30.62
Idaho	42,923	6,120	2,282	8,402	19.57
Illinois	70,813	13,661	3,504	17,165	24.24
Indiana	65,997	11,976	3,253	15,229	23.08
Iowa	69,615	12,782	2,975	15,757	22.63
Kansas	69,171	10,692	3,033	13,725	19.84
Kentucky	76,609	13,817	4,063	17,880	23.34
Louisiana	48,664	8,606	2,931	11,537	23.71
Maine	27,480	4,131	2,934	7,065	25.71
Maryland	35,026	5,465	2,571	8,036	22.94
Massachusetts	25,937	4,040	2,766	6,806	26.24
Michigan	67,207	11,685	3,608	15,293	22.76
Minnesota	66,811	10,502	3,044	13,546	20.28
Mississippi	57,290	10,395	3,448	13,843	24.16
Missouri	79,039	13,874	3,828	17,702	22.40
Montana	46,227	6,863	2,322	9,185	19.87
Nebraska	64,358	10,136	3,000	13,136	20.41
Nevada	23,672	5,369	2,421	7,790	32.91
New Hampshire	21,956	3,352	2,577	5,929	27.00
New Jersey	30,715	5,136	2,509	7,645	24.89
New Mexico	36,864	5,548	2,583	8,131	22.06
New York	54,409	8,636	2,779	11,415	20.98
North Carolina	72,584	12,700	4,031	16,731	23.05
North Dakota	42,442	6,000	2,302	8,302	19.56
Ohio	68,328	11,626	3,429	15,055	22.03
Oklahoma	60,019	8,555	2,807	11,362	18.93
Oregon	43,239	6,122	2,333	8,455	19.55
Pennsylvania	60,220	8,895	3,057	11,952	19.85
Rhode Island	17,055	3,838	2,071	5,909	34.65
South Carolina	46,246	8,234	2,983	11,217	24.26
South Dakota	49,428	7,151	2,274	9,425	19.07
Tennessee	66,694	12,943	3,718	16,661	24.98
Texas	141,618	26,522	7,222	33,744	23.83
Utah	35,200	5,151	2,283	7,434	21.12
Vermont	25,743	3,676	2,509	6,185	24.03
Virginia	69,068	13,190	3,655	16,845	24.39
Washington	44,738	6,361	2,654	9,015	20.15
West Virginia	41,613	7,534	3,345	10,879	26.14
Wisconsin	61,000	8,614	2,721	11,335	18.58
Wyoming	29,424	4,275	2,036	6,311	21.45
Total US and States	**2,556,586**	**430,843**	**151,506**	**582,349**	**22.78**

1/ Includes only non-zero estimates eligible for suppression.

Table 1: Number of Suppressions in the US and State Publications, 2007 Census of Agriculture

3.3 Non-Census Official Estimates Methodology for Confidentiality

For the majority of official estimates generated from NASS sample surveys, protecting the privacy of the survey respondents is not an issue since most estimates are generated at the US or state-level, are very large in magnitude, and are rounded to the number of significant digits the methodology supports. Furthermore, these estimates are generally generated from hundreds, if not thousands of survey responses, each having a sampling and non-response weight. Also, NASS has long tailored its estimation program to generate State-level estimates for only States that produce significant quantities of the given commodity. Hence, the values are large enough that it is not possible to identify individual respondents.

4 NASS Data Laboratory

The current NASS Data Laboratory (commonly referred to as the *Datalab*) was started at the United States Bureau of the Census when it conducted the COA. The Datalab provides two primary services. First, it provides the opportunity for data users to request special tabulations from existing NASS census and survey data. Second, it provides access to micro-level data to approved data users under approved security protocols. Both of these services extend NASS's data products beyond what was anticipated or practicable.

4.1 Special Tabulations

Special tabulation requests may pertain to any NASS census or survey NASS, but most involve COA data, and generally ask for cross tabulations not already provided in the myriad COA publications. The number of special tabulation requests received each year ranges from about 50 to 125. The results of all special tabulations are subjected to the same disclosure avoidance routines used for regular estimates generated from census or sample survey data. Applying the disclosure avoidance routines for special tabulations is complicated by what has already been published – both in NASS's standard publications as well as in previous special tabulations. The complication rises from the necessity to protect suppressed estimates from being deduced from linear combinations of already published tables that were generated at different times and appear in different publications.

4.2 Access to Micro-Data

NASS's Datalab also provides access to micro-level data from censuses and surveys to approved data users under approved security protocols. This access is provided in partnership with USDA's Economic Research Service (ERS) and with NORC at the University of Chicago. Approved researchers have the option to access the micro-level data at either: (1) NASS's headquarters office in Washington, DC, (2) ERS's headquarters office in Washington, DC, (3) one of NASS's 45 Field Offices, or, in the case of ARMS data, (4) the Data Enclave administered by NORC at the University of Chicago. Researches seeking to use NASS census or survey micro-level data must follow a stringent protocol to request access and when utilizing the data. Key elements of this protocol include the following:

- Researchers must complete the NASS form ADM-042, *Request to Access Unpublished Data*. Researchers must include a thorough description of the project, including: (1) the timing of the proposed project, (2) methods of analysis or statistical techniques used, (3) level of reliability required, (4) level of interpretation planned, (5) where the micro-level data will be used, and (6) the specific person(s) who will have access to the data.
- If the request is for access to ARMS micro-level data, a Memorandum of Understanding (MOU) is also required between the institution the researcher is affiliated with, ERS, and NASS.
- Researchers are never provided respondents' identification information.
- Researchers must abide by the same confidentiality policies and laws that NASS staff must abide by. This includes signing a *Confidentiality Certification*, which acknowledges no micro-level data may be released to any other party. It also acknowledges the penalties associated with violating confidentiality: a Class E felony, punishable by up to a $250,000 fine and up to five years in prison.
- Researchers are never permitted to remove micro-level data from the Datalab facility (i.e., NASS or ERS office or NORC Data Enclave). Only summarized output is allowed to be removed, and NASS or ERS staff review all output for confidentiality violations.
- All requests to access NASS micro-level data are reviewed by NASS's Associate Administrator, who also has the sole authority to approve such requests.

5 Conclusion

Maintaining the integrity of confidential data is of paramount concern for all official statistical agencies; disclosing an individual's information is not only unethical, but illegal. However, large-scale censuses and surveys cost taxpayers millions of dollars (or more) so agencies have an obligation to maximize their utility. Fortunately, many techniques exist to substantially reduce disclosure risks to manageable levels. As described in this paper, USDA utilizes multiple data access modalities to disseminate analytically useful data in a secure manner. ARMS data are accessible remotely to authorized researchers through the NORC Data Enclave. Published estimates derived from COA data are made available to the public but only after undergoing a rigorous disclosure control process. In addition, NASS responds to public queries (i.e., special tabulation requests) and releases official estimates generated from sample surveys, protecting the privacy of the survey respondents by the fact that the values are large enough that it is not possible to identify individual respondents. Finally, NASS's Data Laboratory provides the opportunity for data users to request special tabulations from existing NASS census and survey data and provides access to micro-level data to approved data users under approved security protocols.

6 References

Dalenius, T. (1988). Controlled Invasion of Privacy in Surveys, Department of Development and Research, Statistics, Sweden.

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal et al. (Eds): TAMC 2008, LNCS 4978, pp. 1-19.

Fienberg, S.E. and J. McIntyre (2004). "Data Swapping: Variations on a Theme by Dalenius and Reiss," In *Privacy in Statistical Databases: PSD 2004 Proceedings* (J. Domingo-Ferrer and V. Torra, eds.), Lecture Notes in Computer Science, Volume 3050, Springer-Verlag, 14-29.

Lane, J., H. Pascal, and T. Mulcahy (2008). Data Access in a Cyber World: Making Use of Cyberinfrastructure. Transactions on Data Privacy 1, 2-16.

Massell, P. B. (2001). Cell Suppression and Audit Programs Used for Economic Magnitude Data, Statistical Research Report Series No. RR2001/01, United States Bureau of the Census, U.S. Department of Commerce: Washington, DC.

Principles and Practices for a Federal Statistical Agency, Fourth Edition (2009). National Research Council of the National Academies. The National Academies Press: Washington, DC.

Scheuren, F. J. (2009). Presentation on Statistical Matching and other Synthetic and Semi-Synthetic Data Sets: Strengths and Limitations, International Statistical Institute Meetings in Durbin, South Africa.

Skinner, C. (2009). Statistical Disclosure Control for Survey Data. Southampton Statistical Sciences Research Institute.

United Nations, 2007. Managing Statistical Confidentiality *&* Microdata Access: Principles and Guidelines of Good Practice. United Nations Publication Sales No. E.07.II.E.7 ISBN 13: 987-92-1-116959-1 ISSN: 0069-8458.

USDA/NASS Market Research *Report* (2011, forthcoming). Osborn and Barr: St. Louis, MO.

Weinberg, D. H., Abowd, J. M., Rowland, S. K., Steel, P. M., & Zayatz, L. (2007). Access Methods for United States Microdata. Unpublished manuscript,

Washington DC.

Winkler, W. E. (1997). "Producing Public-Use Microdata That are Analytically Valid and Confidential, "Proceedings of the Section on Survey Research Methods, American Statistical Association, 41-50.